RMIT University at TREC 2010: Session Track

Sadegh Kharazmi Falk Scholer Mingfang Wu

School of Computer Science and IT
RMIT University, GPO Box 2476
Melbourne 3001, Australia

1 Introduction

The 2010 session track aimed to investigate retrieval performance over a search session, taking into account the fact that users often need to re-formulate their initial queries to find useful documents.

The experiments carried out by RMIT University investigated a simple strategy of joining query terms across a session, as well as the use of Google suggested queries and whether these can improve the quality of a search result list.

For our experiments, we used the Lemur toolkit (version 4.12) to index and search the ClueWeb category B dataset. Ranking was carried out using a Dirichlet-smoothed language model. Query terms were stemmed using the Krovetz stemmer, and stopwords were not removed. Some queries contained punctuation (for example in URLs), and all punctuation was replaced with whitespace. (The only manual editing was that the the sequence "U.S." was replaced with "USA", but this could have been done automatically through the use of a simple acronym mapping table).

2 Description Of Runs

For the 2010 session track, sessions were composed of two queries: an initial request (RL1) and a paired follow-up request (RL2).

We carried out two experiments using query expansion methods to investigate whether information from RL1 could be used to improve retrieval performance for RL2. The resulting "enhanced" retrieval attempt is labelled RL3.

For the first submission, RMITBase, each query in RL3 is a simple concatenation of the search terms in RL1 and RL2. Duplicate query terms were retained for this run.

Run	nsDCG@10 RL12	nsDCG@10 RL13	nsDCG_dup@10 RL12	nsDCG_dup@10 RL13
RMITBase	0.1374	0.1527	0.1348	0.1454
RMITExp	0.1430	0.1398	0.1450	0.1346
median	0.1945	0.1675	0.1957	0.1759

Table 1: Results based on session-nDCG@10.

Run	nDCG@10 RL1	nDCG@10 RL2	nDCG@10 RL3
RMITBase	0.1245	0.1534	0.1832
RMITExp	0.1283	0.1701	0.1525
median	0.1894	0.1936	0.1501

Table 2: Results based on nDCG@10.

For the second submission, RMITExp, queries were enhanced with "related search" suggestions from Google. Each query in RL1, RL2 and RL3 from RMIT-Base was submitted to Google in turn. All available "search suggestions" were retrieved, and terms were added to the original query. For this run, duplicate query terms were removed from the expanded query.

3 Results

Results for the 2010 session track were evaluated using session-based nDCG at cutoff 10 (nsDCG@10), which accumulates gain values across multiple related queries. An alternative version with duplicate answer items in the system answer lists removed was also generated (nsDCG_dup@10). As can be seen from Table 1, the simple approach of concatenating terms from both queries in the session led to improvements, although the rate of improvement was lower when duplicate answer items were removed. Using query term suggestions harmed results (this may be due to the simple way in which the new terms were added to the original query; it is feasible that appropriate re-weighting could lead to better performance). Both runs performed worse than the median score of all participating systems when evaluated with session-nDCG@10. However, in contrast to the median score, RMITBase led to an improvement in performance when making use of session information.

Table 2 shows results based on nDCG@10, calculated separately for each original query (RL1, RL2) and the enhanced query (RL3). Similar to the previous table, it can be observed that the simple strategy of query concatenation leads to improvements in "session-enhanced" RL3 over what is achieved by using the

Run	D	G	S	Total
RMITBase	20	23	14	57
RMITExp	11	19	10	40

Table 3: Number of queries where the nDCG@10 of RL3 is greater than the nDCG@10 of RL2. Sessions are classified as specification (S), generalization (G) or drifting (D).

"stand-alone" search request RL2. RMITBase shows an improvement in average performance when session information is used, but the difference is not statistically significant (paired t-test, $p = 0.0694$). RMITExp harmed performance, but again not significantly ($p = 0.292$).

The topics for the session track were classified into three categories, based on the type of reformulation that occurred in the session: specification (S), generalization (G) and drifting (D). In Table 3, we show the number of topics for which nDCG@10 for RL3 (the session-enhanced query) outperformed RL2 (the stand-alone query). The concatenation technique used in the RMITBase run was generally effective for both drifting and generalization sessions, but led to fewer improvements for specification sessions. On the other hand, for the RMITExp run, the query suggestion technique appears to have been most effective for generalization sessions, but less useful for drift and specification refinements. However, the differences are not statistically significant for either run ($\chi^2, p > 0.1$).

4 Future Work

When the complete relevance judgements and evaluation scripts become available, we plan to analyse our runs in more detail. In particular, the RL3 queries for the RMITExp run were very long, and we suspect that selective re-weighting of expansion terms might lead to improvements.